Overview of HPC Technologies
Part-I

Hari Subramoni
The Ohio State University
E-mail: subramon@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~subramon

Dhabaleswar K. (DK) Panda
The Ohio State University
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda
HPC: What & Why

• What is High-Performance Computing (HPC)?
  – The use of the most efficient algorithms on computers capable of the highest performance to solve the most demanding problems.

• Why HPC?
  – Large problems – spatially/temporally
    • 10,000 x 10,000 x 10,000 grid → 10^12 grid points → 4x10^12 double variables → 32x10^12 bytes = 32 Tera-Bytes.
    • Usually need to simulate tens of millions of time steps.
    • On-demand/urgent computing; real-time computing;
  – Weather forecasting; protein folding; turbulence simulations/CFD; aerospace structures; Full-body simulation/ Digital human ...

Courtesy: G. Em Karniadakis & L. Grinberg
HPC Examples: Blood Flow in Human Vascular Network

- Cardiovascular disease accounts for about 50% of deaths in western world;
- Formation of arterial disease strongly correlated to blood flow patterns;

**In one minute, the heart pumps the entire blood supply of 5 quarts through 60,000 miles of vessels, that is a quarter of the distance between the moon and the earth**

Computational challenges: Enormous problem size

**Courtesy: G. Em Karniadakis & L. Grinberg**
HPC Examples

Earthquake simulation
Surface velocity 75 sec after earthquake

Flu pandemic simulation
300 million people tracked
Density of infected population, 45 days after breakout

Courtesy: G. Em Karniadakis & L. Grinberg
Trend for Computational Demand

• Continuous increase in demand
  – multiple design choices
  – larger data set
  – finer granularity of computation
  – simulation with finer time step
  – low-latency/high-throughput transaction, …..

• Expectation changes with the availability of better computing systems
Current and Emerging Applications

• High Performance and High Throughput Computing Applications
  – Weather forecasting, physical modeling and simulations (aircraft, engine), drug designs, ...

• Database/Big Data/Machine Learning/Deep Learning applications
  – data-mining, data ware-housing, enterprise computing, machine learning and deep learning

• Financial
  – e-commerce, on-line banking, on-line stock trading

• Digital Library
  – library of audio/video, global library

• Collaborative computing and visualization
  – shared virtual environment

• Telemedicine
  – content-based image retrieval, collaborative visualization/diagnosis

• Virtual Reality, Education and Entertainment
Current and Next Generation Applications and HPC Systems

• Growth of High Performance Computing
  – Growth in processor performance
    • Chip density doubles every 18 months
  – Growth in commodity networking
    • Increase in speed/features + reducing cost

• Clusters: popular choice for HPC
  – Scalability, Modularity and Upgradeability
Integrated High-End Computing Environments

Enterprise Multi-tier Datacenter for Visualization and Mining

Network Based Computing Laboratory 5194.01
Cloud Computing Environments
• Substantial impact on designing and utilizing data management and processing systems in multiple tiers
  – Front-end data accessing and serving (Online)
    • Memcached + DB (e.g. MySQL), HBase
  – Back-end data analytics (Offline)
    • HDFS, MapReduce, Spark
Big Data Analytics with Hadoop

- Underlying Hadoop Distributed File System (HDFS)
- Fault-tolerance by replicating data blocks
- NameNode: stores information on data blocks
- DataNodes: store blocks and host Map-reduce computation
- JobTracker: track jobs and detect failure
- MapReduce (Distributed Computation)
- HBase (Database component)
- Model scales but high amount of communication during intermediate phases
• Three-layer architecture of Web 2.0
  – Web Servers, Memcached Servers, Database Servers
• Memcached is a core component of Web 2.0 architecture
• Distributed Caching Layer
  – Allows to aggregate spare memory from multiple nodes
  – General purpose
• Typically used to cache database queries, results of API calls
• Scalable model, but typical usage very network intensive
Performance Metrics

- **FLOPS, or FLOP/S: **Floating-point Operations Per Second
  - MFLOPS: MegaFLOPS, $10^6$ flops
  - GFLOPS: GigaFLOPS, $10^9$ flops
  - TFLOPS: TeraFLOPS, $10^{12}$ flops
  - PFLOPS: PetaFLOPS, $10^{15}$ flops, present-day supercomputers (www.top500.org)
  - EFLOPS: ExaFLOPS, $10^{18}$ flops, by 2020

- **MIPS**: Million Instructions Per Second

- **What is MIPS rating for iPhone 6?**

  **25,000 MIPS**
  
  **25 GIPS**

*Courtesy: G. Em Karniadakis & L. Grinberg*
High-End Computing (HEC): PetaFlop to ExaFlop

100 PetaFlops in 2017

415 Peta Flops in 2020 (Fugaku in Japan with 7.3M cores)

Expected to have an ExaFlop system in 2021!
Trends for Commodity Computing Clusters in the Top 500 List (http://www.top500.org)

Percentage of Clusters: 94.8%

Number of Clusters:
- Nov-96: 0
- Feb-98: 0
- May-99: 0
- Aug-00: 5
- Nov-01: 10
- Feb-03: 15
- May-04: 20
- Aug-05: 25
- Nov-06: 30
- Feb-08: 35
- May-09: 40
- Aug-10: 45
- Nov-11: 50
- Feb-13: 55
- May-14: 60
- Aug-15: 65
- Nov-16: 70
- Feb-18: 75
- May-19: 80
Drivers of Modern HPC Cluster Architectures

- Multi-core/many-core technologies
- Remote Direct Memory Access (RDMA)-enabled networking (InfiniBand and RoCE)
- Solid State Drives (SSDs), Non-Volatile Random-Access Memory (NVRAM), NVMe-SSD
- Accelerators (NVIDIA GPGPUs and Intel Xeon Phi)
- Available on HPC Clouds, e.g., Amazon EC2, NSF Chameleon, Microsoft Azure, etc.
HPC Technologies

- **Hardware**
  - Interconnects – InfiniBand, RoCE, Omni-Path, etc.
  - Processors – GPUs, Multi-/Many-core CPUs, Tensor Processing Unit (TPU), FPGAs, etc.
  - Storage – NVMe, SSDs, Burst Buffers, etc.

- **Communication Middleware**
  - Message Passing Interface (MPI)
    - CUDA-Aware MPI, Many-core Optimized MPI runtimes (KNL-specific optimizations)
  - NVIDIA NCCL
Major Components in Computing Systems

- Hardware components
  - Processing cores and memory subsystem
  - I/O bus or links
  - Network adapters/switches
- Software components
  - Communication stack

- Bottlenecks can artificially limit the network performance the user perceives
Processing Bottlenecks in Traditional Protocols

- Ex: TCP/IP, UDP/IP
- Generic architecture for all networks
- Host processor handles almost all aspects of communication
  - Data buffering (copies on sender and receiver)
  - Data integrity (checksum)
  - Routing aspects (IP routing)
- Signaling between different layers
  - Hardware interrupt on packet arrival or transmission
  - Software signals between different layers to handle protocol processing in different priority levels
Bottlenecks in Traditional I/O Interfaces and Networks

- Traditionally relied on bus-based technologies (last mile bottleneck)
  - E.g., PCI, PCI-X
  - One bit per wire
  - Performance increase through:
    - Increasing clock speed
    - Increasing bus width
  - Not scalable:
    - Cross talk between bits
    - Skew between wires
    - Signal integrity makes it difficult to increase bus width significantly, especially for high clock speeds

<table>
<thead>
<tr>
<th>I/O Interface</th>
<th>Year</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCI</td>
<td>1990</td>
<td>33MHz/32bit: 1.05Gbps</td>
</tr>
<tr>
<td>PCI-X</td>
<td>1998 (v1.0)</td>
<td>133MHz/64bit: 8.5Gbps</td>
</tr>
<tr>
<td></td>
<td>2003 (v2.0)</td>
<td>266-533MHz/64bit: 17Gbps</td>
</tr>
</tbody>
</table>

Network Based Computing Laboratory
Bottlenecks on Traditional Networks

- Network speeds saturated at around 1Gbps
  - Features provided were limited
  - Commodity networks were not considered scalable enough for very large-scale systems

<table>
<thead>
<tr>
<th>Network Type</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet (1979 -)</td>
<td>10 Mbit/sec</td>
</tr>
<tr>
<td>Fast Ethernet (1993 -)</td>
<td>100 Mbit/sec</td>
</tr>
<tr>
<td>Gigabit Ethernet (1995 -)</td>
<td>1000 Mbit /sec</td>
</tr>
<tr>
<td>ATM (1995 -)</td>
<td>155/622/1024 Mbit/sec</td>
</tr>
<tr>
<td>Myrinet (1993 -)</td>
<td>1 Gbit/sec</td>
</tr>
<tr>
<td>Fibre Channel (1994 -)</td>
<td>1 Gbit/sec</td>
</tr>
</tbody>
</table>
Motivation for InfiniBand and High-speed Ethernet

• Industry Networking Standards
• InfiniBand and High-speed Ethernet were introduced into the market to address these bottlenecks
• InfiniBand aimed at all three bottlenecks (protocol processing, I/O bus, and network speed)
• Ethernet aimed at directly handling the network speed bottleneck and relying on complementary technologies to alleviate the protocol processing and I/O bus bottlenecks
IB Trade Association

- IB Trade Association was formed with seven industry leaders (Compaq, Dell, HP, IBM, Intel, Microsoft, and Sun)
- Goal: To design a scalable and high performance communication and I/O architecture by taking an integrated view of computing, networking, and storage technologies
- Many other industry participated in the effort to define the IB architecture specification
- IB Architecture (Volume 1, Version 1.0) was released to public on Oct 24, 2000
  - Several annexes released after that (RDMA_CM - Sep’06, iSER – Sep’06, XRC – Mar’09, RoCE – Apr’10, RoCEv2 – Sep’14, Virtualization – Nov’16)
  - Latest version 1.3.1 released November 2016
- [http://www.infinibandta.org](http://www.infinibandta.org)
High-speed Ethernet Consortium (10GE/25GE/40GE/50GE/100GE)

- 10GE Alliance formed by several industry leaders to take the Ethernet family to the next speed step
- Goal: To achieve a scalable and high performance communication architecture while maintaining backward compatibility with Ethernet
- http://www.ethernetalliance.org
- 40-Gbps (Servers) and 100-Gbps Ethernet (Backbones, Switches, Routers): IEEE 802.3 WG
- 25-Gbps Ethernet Consortium targeting 25/50Gbps (July 2014)
  - http://25gethernet.org
- Energy-efficient and power-conscious protocols
  - On-the-fly link speed reduction for under-utilized links
- Ethernet Alliance Technology Forum looking forward to 2026
Tackling Communication Bottlenecks with IB and HSE

• Network speed bottlenecks

• Protocol processing bottlenecks

• I/O interface bottlenecks
Network Bottleneck Alleviation: InfiniBand (“Infinite Bandwidth”) and High-speed Ethernet

• Bit serial differential signaling
  – Independent pairs of wires to transmit independent data (called a lane)
  – Scalable to any number of lanes
  – Easy to increase clock speed of lanes (since each lane consists only of a pair of wires)

• Theoretically, no perceived limit on the bandwidth
## Network Speed Acceleration with IB and HSE

<table>
<thead>
<tr>
<th>Network Type</th>
<th>Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet (1979 - )</td>
<td>10 Mbit/sec</td>
</tr>
<tr>
<td>Fast Ethernet (1993 - )</td>
<td>100 Mbit/sec</td>
</tr>
<tr>
<td>Gigabit Ethernet (1995 -)</td>
<td>1000 Mbit/sec</td>
</tr>
<tr>
<td>ATM (1995 -)</td>
<td>155/622/1024 Mbit/sec</td>
</tr>
<tr>
<td>Myrinet (1993 -)</td>
<td>1 Gbit/sec</td>
</tr>
<tr>
<td>Fibre Channel (1994 -)</td>
<td>1 Gbit/sec</td>
</tr>
<tr>
<td>InfiniBand (2001 -)</td>
<td>2 Gbit/sec (1X SDR)</td>
</tr>
<tr>
<td>10-Gigabit Ethernet (2001 -)</td>
<td>10 Gbit/sec</td>
</tr>
<tr>
<td>InfiniBand (2003 -)</td>
<td>8 Gbit/sec (4X SDR)</td>
</tr>
<tr>
<td>InfiniBand (2005 -)</td>
<td>16 Gbit/sec (4X DDR)</td>
</tr>
<tr>
<td>InfiniBand (2007 -)</td>
<td>24 Gbit/sec (12X SDR)</td>
</tr>
<tr>
<td>40-Gigabit Ethernet (2010 -)</td>
<td>40 Gbit/sec</td>
</tr>
<tr>
<td>InfiniBand (2011 -)</td>
<td>54.6 Gbit/sec (4X FDR)</td>
</tr>
<tr>
<td>InfiniBand (2012 -)</td>
<td>2 x 54.6 Gbit/sec (4X Dual-FDR)</td>
</tr>
<tr>
<td>25-/50-Gigabit Ethernet (2014 -)</td>
<td>25/50 Gbit/sec</td>
</tr>
<tr>
<td>100-Gigabit Ethernet (2015 -)</td>
<td>100 Gbit/sec</td>
</tr>
<tr>
<td>Omni-Path (2015 -)</td>
<td>100 Gbit/sec</td>
</tr>
<tr>
<td>InfiniBand (2015 -)</td>
<td>100 Gbit/sec (4X EDR)</td>
</tr>
<tr>
<td>InfiniBand (2017 -)</td>
<td>200 Gbit/sec (4X HDR)</td>
</tr>
</tbody>
</table>

*100 times in the last 16 years*
InfiniBand Link Speed Standardization Roadmap

**InfiniBand Roadmap**

- **XDR** = eXtreme Data Rate
- **NDR** = Next Data Rate
- **HDR** = High Data Rate
- **EDR** = Enhanced Data Rate
- **FDR** = Fourteen Data Rate
- **QDR** = Quad Data Rate
- **DDR** = Double Data Rate (not shown)
- **SDR** = Single Data Rate (not shown)

**link Bandwidth per direction, Gb/s**

- **2008**: 10G
- **2009**: 14G
- **2010**: 25G
- **2011**: 50G
- **2012**: 100G
- **2013**: 200G
- **2014**: 400G
- **2015**: 600G
- **2016**: 1.2T

**Courtesy: InfiniBand Trade Association**
Ethernet Roadmap – To Terabit Speeds?

Terabit speeds by 2025?

50G, 100G, 200G and 400G by 2018-2019

Courtesy: Scott Kipp @ Ethernet Alliance - http://www.ethernetalliance.org/roadmap/
Tackling Communication Bottlenecks with IB and HSE

• Network speed bottlenecks

• Protocol processing bottlenecks

• I/O interface bottlenecks
Capabilities of High-Performance Networks

- Intelligent Network Interface Cards
- Support entire protocol processing completely in hardware (hardware protocol offload engines)
- Provide a rich communication interface to applications
  - *User-level communication capability*
  - Gets rid of intermediate data buffering requirements
- No software signaling between communication layers
  - All layers are implemented on a *dedicated* hardware unit, and not on a *shared* host CPU
Previous High-Performance Network Stacks

- Fast Messages (FM)
  - Developed by UIUC

- Myricom GM
  - Proprietary protocol stack from Myricom

- These network stacks set the trend for high-performance communication requirements
  - Hardware offloaded protocol stack
  - Support for fast and secure user-level access to the protocol stack

- Virtual Interface Architecture (VIA)
  - Standardized by Intel, Compaq, Microsoft
  - Precursor to IB
IB Hardware Acceleration

- Some IB models have multiple hardware accelerators
  - E.g., Mellanox IB adapters

- Protocol Offload Engines
  - Completely implement ISO/OSI layers 2-4 (link layer, network layer and transport layer) in hardware

- Additional hardware supported features also present
  - RDMA, Multicast, QoS, Fault Tolerance, and many more
Ethernet Hardware Acceleration

• Interrupt Coalescing
  – Improves throughput, but degrades latency

• Jumbo Frames
  – No latency impact; Incompatible with existing switches

• Hardware Checksum Engines
  – Checksum performed in hardware $\Rightarrow$ significantly faster
  – Shown to have minimal benefit independently

• Segmentation Offload Engines (a.k.a. Virtual MTU)
  – Host processor “thinks” that the adapter supports large Jumbo frames, but the adapter splits it into regular sized (1500-byte) frames
  – Supported by most HSE products because of its backward compatibility $\Rightarrow$ considered “regular” Ethernet
TOE and iWARP Accelerators

• TCP Offload Engines (TOE)
  – Hardware Acceleration for the entire TCP/IP stack
  – Initially patented by Tehuti Networks
  – Actually refers to the IC on the network adapter that implements TCP/IP
  – In practice, usually referred to as the entire network adapter

• Internet Wide-Area RDMA Protocol (iWARP)
  – Standardized by IETF and the RDMA Consortium
  – Support acceleration features (like IB) for Ethernet

Converged (Enhanced) Ethernet (CEE or CE)

• Also known as “Datacenter Ethernet” or “Lossless Ethernet”
  – Combines a number of optional Ethernet standards into one umbrella as mandatory requirements

• Sample enhancements include:
  – Priority-based flow-control: Link-level flow control for each Class of Service (CoS)
  – Enhanced Transmission Selection (ETS): Bandwidth assignment to each CoS
  – Datacenter Bridging Exchange Protocols (DBX): Congestion notification, Priority classes
  – End-to-end Congestion notification: Per flow congestion control to supplement per link flow control
Tackling Communication Bottlenecks with IB and HSE

- Network speed bottlenecks
- Protocol processing bottlenecks
- I/O interface bottlenecks
Interplay with I/O Technologies

• InfiniBand initially intended to replace I/O bus technologies with networking-like technology
  – That is, bit serial differential signaling
  – With enhancements in I/O technologies that use a similar architecture (HyperTransport, PCI Express), this has become mostly irrelevant now

• Both IB and HSE today come as network adapters that plug into existing I/O technologies
Recent trends in I/O interfaces show that they are nearly matching head-to-head with network speeds (though they still lag a little bit)

<table>
<thead>
<tr>
<th>I/O Interface Type</th>
<th>Year</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCI</td>
<td>1990</td>
<td>33MHz/32bit: 1.05Gbps (shared bidirectional)</td>
</tr>
<tr>
<td>PCI-X</td>
<td>1998 (v1.0) 2003 (v2.0)</td>
<td>133MHz/64bit: 8.5Gbps (shared bidirectional) 266-533MHz/64bit: 17Gbps (shared bidirectional)</td>
</tr>
<tr>
<td>AMD HyperTransport (HT)</td>
<td>2001 (v1.0), 2004 (v2.0) 2006 (v3.0), 2008 (v3.1)</td>
<td>102.4Gbps (v1.0), 179.2Gbps (v2.0) 332.8Gbps (v3.0), 409.6Gbps (v3.1) (32 lanes)</td>
</tr>
<tr>
<td>PCI-Express (PCle) by Intel</td>
<td>2003 (Gen1), 2007 (Gen2), 2009 (Gen3 standard), 2017 (Gen4 standard)</td>
<td>Gen1: 4X (8Gbps), 8X (16Gbps), 16X (32Gbps) Gen2: 4X (16Gbps), 8X (32Gbps), 16X (64Gbps) Gen3: 4X (~32Gbps), 8X (~64Gbps), 16X (~128Gbps) Gen4: 4X (~64Gbps), 8X (~128Gbps), 16X (~256Gbps)</td>
</tr>
<tr>
<td>Intel QuickPath Interconnect (QPI)</td>
<td>2009</td>
<td>153.6-204.8Gbps (20 lanes)</td>
</tr>
</tbody>
</table>
Upcoming I/O Interface Architectures

- Cache Coherence Interconnect for Accelerators (CCIX)
  - https://www.ccixconsortium.com/
- NVLink
- CAPI/OpenCAPI
  - http://opencapi.org/
- GenZ
  - http://genzconsortium.org/
## Comparing InfiniBand with Traditional Networking Stack

### Traditional Ethernet
- **Application Layer**: HTTP, FTP, MPI, File Systems
- **Transport Layer**: Sockets Interface, TCP, UDP
- **Network Layer**: Routing, Flow-control, Error Detection
- **Link Layer**: DNS management tools, Copper, Optical or Wireless
- **Physical Layer**: Copper, Optical or Wireless

### InfiniBand
- **Application Layer**: MPI, PGAS, File Systems
- **Transport Layer**: OpenFabrics Verbs, RC (reliable), UD (unreliable)
- **Network Layer**: OpenSM (management tool)
- **Link Layer**: Flow-control, Error Detection
- **Physical Layer**: Copper or Optical
TCP/IP Stack and IPoIB

Application / Middleware Interface

TCP/IP

Protocol

Kernel Space

Sockets

Adapter

Ethernet Adapter

IPoIB

Switch

Ethernet Switch

InfiniBand Adapter

High Speed Ethernet

InfiniBand Switch

IPoIB
TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Application / Middleware

Verb

RDMA

User Space

InfiniBand Adapter

IB Native

Ethernet Adapter

Switch

Ethernet Switch

High Speed Ethernet

Ethernet Driver

InfiniBand Adapter

Switch

InfiniBand Switch

IPoIB

IB Native

TCP/IP, IPoIB and Native IB Verbs

Applica
Components: Channel Adapters

- **Used by processing and I/O units to connect to fabric**
- **Consume & generate IB packets**
- **Programmable DMA engines with protection features**
  - Independent buffering channeled through Virtual Lanes
- **May have multiple ports**
- **Host Channel Adapters (HCAs)**
Components: Switches and Routers

- Relay packets from a link to another
- Switches: intra-subnet
- Routers: inter-subnet
- May support multicast
Components: Links & Repeaters

- Network Links
  - Copper, Optical, Printed Circuit wiring on Back Plane
  - Not directly addressable

- Traditional adapters built for copper cabling
  - Restricted by cable length (signal integrity)
  - For example, QDR copper cables are restricted to 7m

- Intel Connects: Optical cables with Copper-to-optical conversion hubs (acquired by Emcore)
  - Up to 100m length
  - 550 picoseconds copper-to-optical conversion latency

- Available from other vendors (Luxtera)
- Repeaters (Vol. 2 of InfiniBand specification)
Communication in the Channel Semantics (Send/Receive Model)

Send WQE contains information about the send buffer (multiple non-contiguous segments)

Receive WQE contains information on the receive buffer (multiple non-contiguous segments); Incoming messages have to be matched to a receive WQE to know where to place

Processor is involved only to:
1. Post receive WQE
2. Post send WQE
3. Pull out completed CQEs from the CQ
Communication in the Memory Semantics (RDMA Model)

Initiator processor is involved only to:
1. Post send WQE
2. Pull out completed CQE from the send CQ

No involvement from the target processor

Send WQE contains information about the send buffer (multiple segments) and the receive buffer (single segment)
Communication in the Memory Semantics (Atomics)

Send WQE contains information about the send buffer (single 64-bit segment) and the receive buffer (single 64-bit segment)

Initiator processor is involved only to:
1. Post send WQE
2. Pull out completed CQE from the send CQ

No involvement from the target processor

IB supports compare-and-swap and fetch-and-add atomic operations
IB Multicast Example

Switch decodes inbound packet header (LRH) DLID to determine target output ports.

Router decodes inbound packet header (GRH) GID multicast address to determine target output ports.
IPoIB vs. SDP Architectural Models

**Traditional Model**
- Sockets Application
- Sockets API

**Possible SDP Model**
- Sockets Application
- Sockets API

- Kernel
  - TCP/IP Sockets Provider
  - TCP/IP Transport Driver
  - IPoIB Driver

- User

- Kernel
  - TCP/IP Sockets Provider
  - TCP/IP Transport Driver
  - IPoIB Driver

- User

*InfiniBand CA*

*SDP*

*OS Modules*

*InfiniBand Hardware*

(Source: InfiniBand Trade Association)
RSockets Overview

- Implements various socket like functions
  - Functions take same parameters as sockets
- Can switch between regular Sockets and RSockets using LD_PRELOAD
TCP/IP, IPoIB, Native IB Verbs, SDP and RSocket
tics

- Application / Middleware Interface
- Sockets
- Verbs

- Protocol
  - Kernel Space
    - TCP/IP
    - IPoIB
    - Ethernet Driver
    - InfiniBand Adapter
    - Ethernet Adapter
    - InfiniBand Switch
    - Ethernet Switch
    - High Speed Ethernet

- Adapter
  - User Space
    - RSocket
    - InfiniBand Adapter
    - IPoIB
    - InfiniBand Switch
    - IPoIB

- Switch
  - User Space
  - InfiniBand Adapter
  - IPoIB
  - InfiniBand Switch
  - IPoIB

Network Based Computing Laboratory
5194.01
53
RDMA over Converged Enhanced Ethernet (RoCE)

**Network Stack Comparison**

- **Software** written with IB-Verbs
- **Link layer** is Converged (Enhanced) Ethernet (CE)
- **100Gb/s support** from latest EDR and ConnectX-3 Pro adapters

**Pros: IB Vs RoCE**
- Works natively in Ethernet environments
  - Entire Ethernet management ecosystem is available
- Has all the benefits of IB verbs
- Link layer is very similar to the link layer of native IB, so there are no missing features

**RoCE v2: Additional Benefits over RoCE**
- Traditional Network Management Tools Apply
- ACLs (Metering, Accounting, Firewalling)
- GMP Snooping for Optimized Multicast
- Network Monitoring Tools

---

**Packet Header Comparison**

- ETH L2 Hdr
- IB GRH L3 Hdr
- IB BTH+ L4 Hdr

- ETH L2 Hdr
- IP Hdr L3 Hdr
- UDP Hdr
- Port #
- IB BTH+ L4 Hdr

---

**Courtesy: OFED, Mellanox**
A Brief History of Omni-Path

- Pathscale (2003 – 2006) came up with initial version of IB-based product
- QLogic enhanced the product with the PSM software interface
- The IB product of QLogic was acquired by Intel
- Intel enhanced the QLogic IB product to create the Omni-Path product
Omni-Path Fabric Overview

• Layer 1.5: Link Transfer Protocol
  – Features
    • Traffic Flow Optimization
    • Packet Integrity Protection
    • Dynamic Lane Switching
  – Error detection/replay occurs in Link Transfer Packet units
  – 1 Flit = 65b; LTP = 1056b = 16 flits + 14b CRC + 2b Credit
  – LTPs implicitly acknowledged
  – Retransmit request via NULL LTP; carries replay command flit

• Layer 2: Link Layer
  – Supports 24 bit fabric addresses
  – Allows 10KB of L4 payload; 10,368 byte max packet size
  – Congestion Management
    • Adaptive / Dispersive Routing
    • Explicit Congestion Notification
  – QoS support
    • Traffic Class, Service Level, Service Channel and Virtual Lane

• Layer 3: Data Link Layer
  – Fabric addressing, switching, resource allocation and partitioning support
## IB, Omni-Path, and HSE: Feature Comparison

<table>
<thead>
<tr>
<th>Features</th>
<th>IB</th>
<th>iWARP/HSE</th>
<th>RoCE</th>
<th>RoCE v2</th>
<th>Omni-Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware Acceleration</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>RDMA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Congestion Control</td>
<td>Yes</td>
<td>Optional</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Multipathing</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Atomic Operations</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Multicast</td>
<td>Optional</td>
<td>No</td>
<td>Optional</td>
<td>Optional</td>
<td>Optional</td>
</tr>
<tr>
<td>Data Placement</td>
<td>Ordered</td>
<td>Out-of-order</td>
<td>Ordered</td>
<td>Ordered</td>
<td>Ordered</td>
</tr>
<tr>
<td>Prioritization</td>
<td>Optional</td>
<td>Optional</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Fixed BW QoS (ETS)</td>
<td>No</td>
<td>Optional</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Ethernet Compatibility</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>TCP/IP Compatibility</td>
<td>Yes (using IPoIB)</td>
<td>Yes (using IPoIB)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Software Convergence with OpenFabrics

• Open source organization (formerly OpenIB)
  – [www.openfabrics.org](http://www.openfabrics.org)

• Incorporates both IB, RoCE, and iWARP in a unified manner
  – Support for Linux and Windows

• Users can download the entire stack and run
  – Latest stable release is OFED 4.8
    • New naming convention to get aligned with Linux Kernel Development
    • OFED 4.8.1 is under development
OpenFabrics Stack with Unified Verbs Interface

Verbs Interface
(libibverbs)

User Level

Mellanox
(libmthca, libmlx*)
Intel/QLogic
(libipathverbs)
IBM
(libehca)
Chelsio
(libcxgb*)
Emulex
(libocrdma)
Intel/NetEffect
(libnes)

Kernel Level

Mellanox
(ib_mthca, ib_mlx*)
Intel/QLogic
(ib_ipath)
IBM
(ib_ehca)
Chelsio
(ib_cxgb*)
Emulex
Adapters
Intel/NetEffect
Adapters

Mellanox
Adapters
Intel/QLogic
Adapters
IBM
Adapters
Chelsio
Adapters
Emulex
Adapters
Intel/NetEffect
Adapters
OpenFabrics Software Stack

Application Level
- User Level MAD API
  - InfiniBand
  - IP Based App Access
  - Sockets Based Access
  - Various MPIs
  - Block Storage Access
  - Clustered DB Access
  - Access to File Systems

User APIs
- User Level MAD API
  - OpenFabrics User Level Verbs / API
  - iWARP R-NIC

Upper Layer Protocol
- User Space
  - User APIs
  - IPoIB
  - SDP Lib
  - SRP
  - iSER
  - RDS
  - NFS-RDMA
  - Cluster File Sys

Mid-Layer
- Kernel Space
  - Connection Manager Abstraction (CMA)
  - Connection Manager
  - InfiniBand
  - OpenFabrics Kernel Level Verbs / API
  - iWARP R-NIC

Provider
- Hardware Specific Driver
  - InfiniBand HCA

Key
- Common
- InfiniBand
- iWARP

Network Based Computing Laboratory
Libfabrics Software Stack

Open Fabrics Interface (OFI)
- Control Services
  - Discovery
- Communication Services
  - Connection Management
  - Address Vectors
- Completion Services
  - Event Queues
  - Counters
- Data Transfer Services
  - Message Queues
  - RMA
  - Tag Matching
  - Atomics
  - Triggered Operations

OFI Provider
- Discovery
- Connection Management
- Address Vectors
- Event Queues
- Counters
- Message Queues
- RMA
- Tag Matching
- Atomics
- Triggered Operations

NIC
- TX Command Queues
- RX Command Queues

OFI Enabled Applications

## Large-scale InfiniBand Installations

- 153 IB Clusters (30.6%) in the June’20 Top500 list
  - (http://www.top500.org)
- Installations in the Top 50 (19 systems):

<table>
<thead>
<tr>
<th>System Description</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2,414,592 core (Summit) at ORNL (2nd)</td>
<td>127,488 core (DGX SuperPOD) at NVIDIA (23rd)</td>
</tr>
<tr>
<td>1,572,480 core (Sierra) at LLNL (3rd)</td>
<td>204,032 core (Gadi) at Fujitsu/Lenovo (24th)</td>
</tr>
<tr>
<td>669,670 core (HPC5) at Italy (6th)</td>
<td>170,352 core (Taiwania 2) in Taiwan (25th)</td>
</tr>
<tr>
<td>272,800 core (Selene) at NVIDIA (7th)</td>
<td>130,000 core (AiMOS) in IBM (26th)</td>
</tr>
<tr>
<td>448,448 core (Frontera) at TACC (8th)</td>
<td>174,720 core (Roxy) at HPE (28th)</td>
</tr>
<tr>
<td>347,776 core (Marconi-100) in Italy (9th)</td>
<td>294,912 core (Blenos) at Atos (29th)</td>
</tr>
<tr>
<td>391,680 core (ABCI) at AIST/Japan (12th)</td>
<td>169,920 core (PupMaya) in US (30th)</td>
</tr>
<tr>
<td>288,288 core (Lassen) at LLNL (14th)</td>
<td>107,568 core (Artemis) in NVIDIA (31st)</td>
</tr>
<tr>
<td>291,024 core (Pangea III) in France (15th)</td>
<td>197,120 core (JOLIOT-CURIE ROME) at CEA/France (33rd)</td>
</tr>
<tr>
<td>253,600 core (HPC4) in Italy (19th)</td>
<td>and many more!</td>
</tr>
</tbody>
</table>
Large-scale Omni-Path Installations

- 49 Omni-Path Clusters (9.8%) in the June’20 Top500 list
  - [http://www.top500.org](http://www.top500.org)
- Installations in the Top 50 (19 systems):

<table>
<thead>
<tr>
<th>Core Count</th>
<th>Location/Provider</th>
</tr>
</thead>
<tbody>
<tr>
<td>305,586 core</td>
<td>SuperMUC-NG in Germany (13th)</td>
</tr>
<tr>
<td>570,020 core</td>
<td>Nurion at KISTI/South Korea (17th)</td>
</tr>
<tr>
<td>556,104 core</td>
<td>Oakforest-PACS at JCAHPC in Japan (18th)</td>
</tr>
<tr>
<td>367,024 core</td>
<td>Stampede2 at TACC in USA (21st)</td>
</tr>
<tr>
<td>348,000 core</td>
<td>Marconi XeonPhi at CINECA in Italy (22nd)</td>
</tr>
<tr>
<td>135,828 core</td>
<td>Tsubame 3.0 at TiTech in Japan (27th)</td>
</tr>
<tr>
<td>153,216 core</td>
<td>MareNostrum at BSC in Spain (37th)</td>
</tr>
<tr>
<td>127,520 core</td>
<td>Cobra in Germany (44th)</td>
</tr>
<tr>
<td>103,680 core</td>
<td>Lise in Germany (48th)</td>
</tr>
<tr>
<td>86,800 core</td>
<td>TX-GAIA in MIT Lincoln Labs (50th)</td>
</tr>
<tr>
<td>93,960 core</td>
<td>Jean Zay in France (54th)</td>
</tr>
<tr>
<td>76,608 core</td>
<td>Oakforest-CX in Japan (59th)</td>
</tr>
<tr>
<td>86,400 core</td>
<td>Joule 2.0 at NETL/DOE/USA (69th)</td>
</tr>
<tr>
<td>67,584 core</td>
<td>Cedar (GPU) in Canada (74th)</td>
</tr>
<tr>
<td>62,400 core</td>
<td>LLNL/NNSA CTS-1 MAGMA in USA (79th)</td>
</tr>
<tr>
<td>55,296 core</td>
<td>Mustang at AFRL/USA (80th)</td>
</tr>
<tr>
<td>53,568 core</td>
<td>Numerical Material Simulator in Japan (87th)</td>
</tr>
<tr>
<td>61,120 core</td>
<td>Jean Zay in France (91st)</td>
</tr>
<tr>
<td>64,512 core</td>
<td>Tetralith at NSC/Sweden (95th)</td>
</tr>
</tbody>
</table>

*and many more!*
Ethernet-based Scientific Computing Installations

- 263 Ethernet-based (1G, 10G, 25G, 50G, 100G, 200G) compute systems with ranking in the June’20 Top500 list
  - 163,840-core 200G installation in China (#57)
  - 76,000-core 25G installation in China (#86) – new
  - 73,600-core 25G installation in China (#94) – new
  - 72,000-core 25G installation in China (#96) – new
  - 68,000-core 25G installation in China (#102) – new
  - 66,400-core 25G installation in China (#103) – new
  - 128,000-core 25G installation in China (#109) – new
  - 64,000-core 25G installation in China (#112) – new
  - 62,400-core 25G installation in China (#117) – new
  - 64,320-core 25G installation in China (#118) – new
  - 61,440-core 25G installation in China (#125) – new
  - 50,400-core 25G installation in China (#128) – new
  - 50,400-core 100G installation in China (#129) – new
  - 60,000-core 25G installation in China (#131) – new
  - 59,200-core 25G installation in China (#134) – new
  - 115,200-core 10G installation in China (#135) – new
  - 58,800-core 25G installation in China (#136) – new
  - 113,600-core 25G installation in China (#137) – new
  - and many more!